УДК 004.93’1

С.О. Медведєва
С.С. Давидюк

OPTICAL CHARACTER RECOGNITION

Вінницький національний технічний університет

Анотація

Проведено загальний аналіз технології оптичного розпізнавання символів. Розглянуто основні етапи становлення, проведено порівняння двох базових алгоритмів оптичного розпізнавання символів та відзначено переваги застосування і перспективи розвитку даної технології.

Ключові слова

Оптичне розпізнавання символів, машинно-закодований текст, комп'ютерний зір, телеграфний код, комп'ютерна програма,система перебору спаму, КАПЧА, WebOCR, алгоритм оптичного розпізнавання символів, зіставлення зі зразком, кореляція зображень, синтезатори мови.

Abstract

The work analyses the thechnology of optical character recognition. It reveals the main stages of its development, compares two basic types of core OCR algorithms, underlines the benefits of OCR application and shows the prospects of its development.

Keywords

Optical character recognition, machine-encoded text, computer vision, telegraph code, computer program, spam-busting system, CAPTCHA, WebOCR, OCR algorithm, pattern matching, image correlation, speech synthesizers.

Character Recognition (CR) has been extensively studied in the last half century and progressed to a level, sufficient to produce technology driven applications. Now, the rapidly growing computational power enables the implementation of the present CR methodologies and also creates an increasing demand on many emerging application domains, which require more advanced methodologies.

Optical character recognition (OCR) is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text.

Early versions needed to be trained with images of each character, and could work on one font at a time. Advanced systems capable of producing a high degree of recognition accuracy for most fonts are now common. OCR is a field of research in pattern recognition, artificial intelligence and computer vision.

Early optical character recognition can be traced back to technologies involving telegraphy and creating reading devices for the blind. In 1914, Emanuel Goldberg developed a machine that read characters and converted them into standard telegraph code.

In 1974, Ray Kurzweil started the company Kurzweil Computer Products, Inc. and continued the development of omni-font OCR, which could recognize text printed in virtually any font. Kurzweil decided that the best application of this technology would be to create a reading machine for the blind, which would allow blind people to have a computer read text to them loudly. In 1978, Kurzweil Computer Products began selling a commercial version of the optical character recognition computer program.

In the 2000s Researchers at Carnegie Mellon University developed a spam-busting system called CAPTCHA.

OCR became available online as a service (WebOCR), in a cloud computing environment, and in mobile applications like real-time translation of foreign-language signs on a smartphone.

Various commercial and open source OCR systems are nowadays available for most common writing systems.

All OCR systems include an optical scanner for reading text, and sophisticated software for analyzing images. Most OCR systems use a combination of hardware (specialized circuit boards) and software to recognize characters.

Here are two basic types of core OCR algorithm, which can produce a ranked list of candidate characters.

Matrix matching involves comparing an image to a stored glyph on a pixel-by-pixel basis; it is also known as "pattern matching", "pattern recognition", or "image correlation". This technique works best with typewritten text and does not work well when new fonts are encountered.

Feature extraction decomposes glyphs into "features" like lines, closed loops, line direction, and line intersections. These are compared with an abstract vector-like representation of a character, which might reduce to one or more glyph prototypes. General techniques of feature detection in computer vision are applicable to this type of OCR, which is commonly seen in "intelligent" handwriting recognition and indeed most modern OCR software.

Once a printed page is in this machine-readable text form, we can search through it by keyword, edit it with a word processor, incorporate it into a Web page, compress it into a ZIP file and store it in much less space, send it by email — and all kinds of other useful things. Machine-readable text can also be decoded by screen readers, tools that use speech synthesizers to read out the words on a screen so blind and visually impaired people can understand them.

The potential of OCR systems is enormous because they enable users to harness the power of computers to access printed documents. OCR is already being used widely in the legal profession, where searches that once required hours or days can now be accomplished in a few seconds.

Through the years, the methods of character recognition have improved from quite primitive schemes, suitable only for reading stylized printed numerals, to more complex and sophisticated techniques for the recognition of a great variety of typeset fonts and also handprinted characters.

Generally there is a potential in using context to a larger extent than what is done today. In addition, combinations of multiple independent feature sets and classifiers, where the weakness of one method is compensated by the strength of another, can improve the recognition of individual characters.

The frontiers of research within character recognition have now moved towards the recognition of cursive script, which is handwritten connected or calligraphic characters. Promising techniques within this area, deal with the recognition of entire words instead of individual characters.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ / REFERENCES

1. Schantz, Herbert F. The history of OCR, optical character recognition. [Manchester Center, Vt.]: Recognition Technologies Users Association, 1982. ISBN 9780943072012.

2. George Nagy. Optical Character Recognition: An illustrated guide to the frontier. George Nagy, Thomas A. Nartker, Stephen V. Rice. Procs. Document Recognition and Retrieval VII, SPIE Vol. 3967, 58-69.

3. Mohamed Cheriet. Character recognition systems : a guide for students and practioners / Mohamed Cheriet ... [et al.]. p. cm. ISBN9780471415701.

4. Pati, P.B. Word Level Multi-script Identification. Pattern Recognition Letters, Pati, P.B.; Ramakrishnan, A.G – Vol. 29, pp. 1218 - 1229, 1987.

Науковий керівник: Медведєва Світлана Олександрівна – викладач кафедри іноземних мов, Вінницький національний технічний університет.

Давидюк Світлана Сергіївна – студентка групи 1Сі- 14б, факультет комп'ютерних систем та автоматики, Вінницький національний технічний університет, м. Вінниця, svitlanadavydiuk@gmail.com.

Supervisor: Svitlana Medvedieva – teacher of English, the Foreign Languages Department, Vinnytsia National Technical University, Vinnytsia

Svitlana Davydiuk– student, group 1SI-12, Faculty of Computer Systems and Automatics, Vinnytsia National Technical University, Vinnytsia